Leidos-SERAS
2890 Woodbridge Avenue, Building 209 Annex
Edison, NJ 08837-3679
Telephone: 732-321-4200, Facsimile: 732-494-4021

DATE: March 15, 2018

TO: Felicia Barnett, Director SCMTSC, EPA WAM

FROM: Donna J. Getty, SERAS Statistician

THROUGH: Richard Leuser, SERAS Deputy Program Manager, Task Leader

SUBJECT: FINAL -- STATISTICAL REVIEW OF UNITED STATES NAVY PROPOSED WORK PLAN FOR RADIOLOGICAL SURVEY AND SAMPLING, HUNTERS POINT NAVAL SHIPYARD SITE (HPNS), SERAS-106, WORK ORDER #83

INTRODUCTION

United States (US) Environmental Protection Agency (EPA) Region 9 personnel requested a statistical review of the sampling strategies outlined in the US Navy's proposed *Work Plan, Radiological Survey and Sampling, Former Hunters Point Naval Shipyard, San Francisco, California* (February 2018) (Work Plan). The Work Plan presents an overview of proposed methodologies and protocols for the investigation of radiological contamination in Parcels B, C, D-2, E, G, UC-1, UC-2, and UC-3 on the Hunters Point Naval Shipyard (HPNS) (Site), located in San Francisco, California. Allegations of Navy subcontractor data falsification in reference to historical remedial work conducted at the Site has resulted in the need for an assessment of potential residual radiological contamination potentially left on-site. The Navy has proposed to perform characterization surveys and final status surveys (FSS), and then remediate as needed. The US EPA's objective is to assess the full extent of radiological contamination which remains at these Parcels on HPNS for the identified radionuclides of concern (ROCs), those present in background and those not present in background, in: surface soils (land areas), within previously defined trench units (the backfill, walls, bottom and immediate peripheral soil), and buildings.

This Technical Memorandum (TM) formalizes comments which were made verbally and via emails to the EPA Region 9 Remedial Project Manager (RPM) and the HPNS team of Regulators. Specifically, comments on Sections:

4.1.2.1 Soil Investigation Levels
4.2 Reference Backgrounds
4.3 Soil Survey Areas
4.4 Building Survey areas
4.5 Data Quality Objectives
5.2 Surface and Subsurface Soil Investigations
5.3 Former Building and Pavement Investigations
5.4 Building Investigations
6 Data Evaluation

Regulator concerns include proposed sample sizes for background and site areas, statistical testing, confidence levels, interpretation of final results, and application of *Multi-Agency Radiation Survey and Site Investigation Manual* (MARSSIM, 2002) and *Multi-Agency Radiation Survey and Assessment of Materials and Equipment* (MARSAME, 2009) strategies.

GENERAL COMMENTS

Comment 1: Application of MARSSIM. In reviewing the Work Plan, proposed strategies/methodologies were compared with recommended strategies from MARSSIM (2002) which is frequently referenced in the Work Plan. However, it is important to recognize that MARSSIM does not provide guidance on sampling strategies for subsurface soil contamination. It specifically addresses surface contamination in land areas and buildings. As stated on page 5-51 of MARSSIM:

"In addition to the building and land surface areas described above, there are numerous other locations where measurements and/or sampling may be necessary. Examples include items of equipment and furnishings, building fixtures, drains, ducts, and piping. Many of these items or locations have both internal and external surfaces with potential residual radioactivity. Subsurface measurements and/or sampling may also be necessary. Guidance on conducting or evaluating these types of surveys is outside the scope of MARSSIM."

All subsurface sampling strategies presented in the HPNS Work Plan are outside of the scope of MARSSIM. However, many of the statistical methodologies presented in MARSSIM can be adapted to subsurface soils if appropriate sampling protocols and relevant statistical methodologies are applied. All proposed methodologies for subsurface soil evaluation were reviewed for statistical validity and to determine the adequacy of the proposed sample sizes.

Comment 2: Application of MARSAME. The MARSAME manual supplements MARSSIM and provides technical information on survey approaches to determine proper disposition of materials and equipment (M&E). Guidance within this manual was also reviewed to assess its application to the HPNS Site. Similar to MARSSIM, MARSAME does not specifically address subsurface soils:

"The scope of MARSAME is M&E potentially affected by radioactivity, including metals, concrete, tools, equipment, piping, conduit, furniture and dispersible bulk materials such as trash, rubble, roofing materials, and sludge." (MARSAME, pg. RM-1)

"Examples of M&E include metals, concrete, tools, equipment, piping, conduit, furniture, and dispersible bulk materials such as trash, rubble, roofing materials, and sludge. Liquids, gases, and solids stored in containers (e.g., drums of liquid, pressurized gas cylinders, containerized soil) are also included in the scope of this document." (MARSAME, pg. 1-1)

Like MARSSIM, statistical analyses presented within MARSAME can be adapted for evaluation of subsurface soils if assumptions associated with the statistical analyses are met and adequate sample sizes are computed.

SPECIFIC COMMENTS

Comment 3: Section 4.1.2.1 Soil Investigation Levels – Second Paragraph, page 4-12

"*The investigation level for gamma scan results will be established at three standard deviations above the mean for the gamma scan data set being evaluated."*

Reviewer Comments: As read, this implies that the Navy will determine an investigation level (IL), for each survey scan they conduct, based on the mean of the data they collect during that scan. As proposed in the Work Plan, survey scans will be conducted per defined sample unit (SU). If the Navy uses the mean per scan survey, it can lead to higher ILs and less recognized contamination.

Gamma scan data is measured as count data not continuous data. It is well established that count data typically follow what is called a Poisson distribution as opposed to a normal distribution (Gaussian curve). The variance of a Poisson distribution is equal to the mean. This implies that as the mean of the survey scan data increases, the standard deviation (square root of the variance) increases, hence the IL increases (3 standard deviations above the mean). When large numbers of count data are collected the distribution approximates a Gaussian curve, but still retains the property that the mean is approximately equal to the variance.

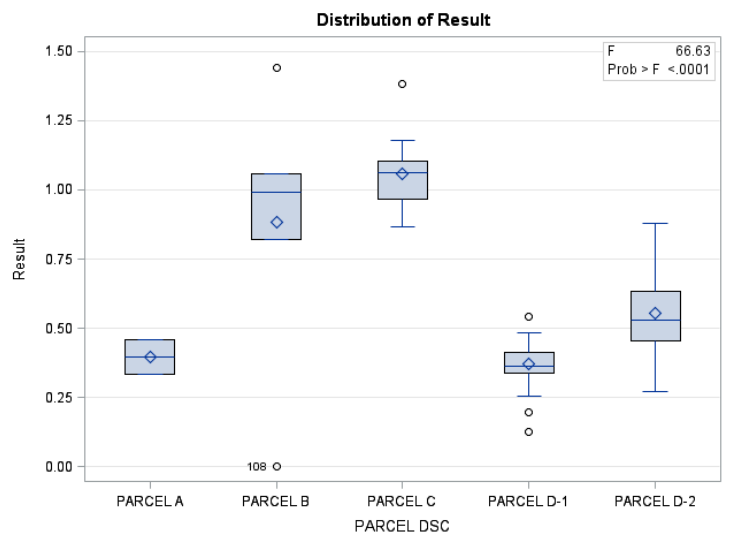
It is recommended that the IL for ROCs found in background should be based on background reference area measurements with similar soil type to the SU being evaluated, to ensure identification of residual contamination.

Comment 4: Section 4.2.1 Soil Reference Areas – page 4-14

Navy proposes five surface and 25 subsurface samples collected at a minimum of five background reference areas to establish concentrations of Ra-226 and Cs-137 in soils, cites MARSSIM guidance as requiring a minimum of 18 measurements per SU and each background area and the Nuclear Regulatory Commission (NRC) as requiring at least 100 samples from at least 5 distinct locations. The Navy has proposed increasing the minimum requirement of 18 measurements to 25 to ensure sufficient background data would be available.

Reviewer Comments: In order to meet MARSSIM and NRC criteria, sample sizes for background reference areas should be computed per independent interval, surface soils and subsurface soils, not across both. Additionally, it is clear from historical data provided by the Navy and collected at the background reference areas in Parcels A, B, C, D-1, and D-2, that variability is not consistent across the five areas for off-site laboratory Ra-226 measurements (Figure 1).

Figure 1. Distribution of Ra-226 Off-Site Laboratory Measurements in picocuries per gram (pCi/g) in Background Reference Areas



Recommendations: Given the differences in variability and mean/median concentrations for Ra-226 as demonstrated in Figure 1, it is recommended that background reference area sample data should not be combined across the five areas, but rather background reference areas should be established per Parcel with sample sizes computed based on the variability within each background reference area per independent interval (surface soils and subsurface soils).

Sample sizes should be justified with detailed statistical analyses and explanations of how inputs to the computations were derived, including specifics of how estimates of variability were obtained (e.g., what data was included in the calculation, how and where the data was collected, what assumptions were made). If measurements from multiple background reference areas will be combined, the results of a comparative analysis such as an Analysis of Variance (ANOVA) or non-parametric equivalent (Kruskal-Wallis Test) must be documented to support combining the areas. These comparative tests will establish if there is a statistically significant difference between Ra-226 and Cs-137 in the background reference areas at a specified confidence-level.

Comment 5: Section 4.3.3 Number of Samples in a Survey Unit – page 4-19

Table 4-3 provides the inputs the Navy used to compute the number of samples which will be collected from each SU based on Section 5.5.2.1 of MARSSIM. Inputs include the $DCGL_w$, Δ ($DCGL_w$ – Background), σ (standard deviation),

α (probability of rejecting the null hypothesis when it is true), and β (probability of accepting the null hypothesis [H_0] when it is false).

Reviewer Comments: Proposed sample sizes only apply to ROCs which have been identified in background, Ra-226 and Cs-137, and are representative of the number of samples which need to be collected from each SU and background reference area to achieve the chosen confidence levels (α=0.01 and β=0.01) when running a Wilcoxon Rank-Sum (WRS) test, if the proposed estimates of σ are valid.

The reviewer agrees with the chosen confidence level for α and β of 0.01 as it is the most conservative and protective of human health.

The Work Plan uses σ =0.28 for Ra-226 and σ=0.113 as inputs to the sample size calculations. To evaluate the validity of these proposed estimates of σ, basic descriptive statistics including standard deviation were computed for historical background reference measurements for Ra-226 and Cs-137 (Tables 1 and 2). Table 1 provides the descriptive statistics for the historical off-site laboratory measurements and Table 2 for the historical on-site laboratory measurements. The statistics are computed per nuclide and parcel. For off-site laboratory measurements, Cs-137 σ ranges from 0 in Parcel A to 0.0498 in Parcel B and Ra-226 σ ranges from 0.0788 in Parcel D-1 to 0.479 in Parcel B. For on-site laboratory measurements, Cs-137 σ ranges from 0.0310 in Parcel D-1 to 0.0456 in Parcel B and Ra-226 σ ranges from 0.274 in Parcel D-2 to 0.471 in Parcel B.

Note that variability is greatest for both Cs-137 and Ra-226 in Parcel B for on-site and off-site laboratory measurements. This supports the recommendation that background reference areas should be established per Parcel with sample sizes computed per background reference area. If consistency is preferred, then sample sizes across Parcels for on-site SUs and background reference areas should be based on the reference background area with the greatest variability.

Table 1. Original Background Data/FRED - Sigma (σ) for Cs-137 and Ra-226
By Parcel/Site for Off-Site Laboratory Results

Nuclide=Cs-137

Analysis Variable : Result						
PARCEL	**SITEDSC**	**N Obs**	**N**	**Mean**	**Median**	**Std Dev (σ)**
PARCEL A	**Building 901**	**2**	2	0	0	0
PARCEL B	**Building 116**	**6**	6	0.0150000	-0.0015000	0.0497835
PARCEL C	**TURAC**	**18**	18	-0.0019527	-0.0027005	0.0103421
PARCEL D-1	**Building 526 Berth 29**	**40**	40	0.0022800	0	0.0125294
PARCEL D-2	**Building 813 Lot**	**36**	36	0.000254861	0.000055500	0.0107954

Nuclide=Ra-226

Analysis Variable : Result						
PARCEL	**SITEDSC**	**N Obs**	**N**	**Mean**	**Median**	**Std Dev (σ)**
PARCEL A	**Building 901**	**2**	2	0.3965000	0.3965000	0.0883883
PARCEL B	**Building 116**	**6**	6	0.8836667	0.9900000	0.4793085
PARCEL C	**TURAC**	**18**	18	1.0572611	1.0610000	0.1176851
PARCEL D-1	**Building 526 Berth 29**	**40**	40	0.3703000	0.3635000	0.0787987
PARCEL D-2	**Building 813 Lot**	**36**	36	0.5562500	0.5280000	0.1443607

Table 2. Original Background Data/FRED - Sigma (σ) for Cs-137 and Ra-226
By Parcel/Site for On-Site Laboratory Results

Nuclide=Cs-137

Analysis Variable : Result						
PARCEL	**SITEDSC**	**N Obs**	**N**	**Mean**	**Median**	**Std Dev (σ)**
PARCEL A	**Building 901**	**18**	18	0.0337060	0.0266382	0.0359294
PARCEL B	**Building 116**	**37**	37	0.0240768	0.0247570	0.0455666
PARCEL D-1	**Building 526 Berth 29**	**20**	20	0.0368457	0.0293925	0.0310762
PARCEL D-2	**Building 813 Lot**	**18**	18	-0.0218994	-0.0215520	0.0380905

Nuclide=Ra-226

Analysis Variable : Result						
PARCEL	**SITEDSC**	**N Obs**	**N**	**Mean**	**Median**	**Std Dev (σ)**
PARCEL A	**Building 901**	**18**	18	0.3626028	0.1038700	0.4403894
PARCEL B	**Building 116**	**37**	37	0.4477704	0.3972000	0.4713866
PARCEL D-1	**Building 526 Berth 29**	**20**	20	0.6331562	0.6552700	0.3061107
PARCEL D-2	**Building 813 Lot**	**18**	18	0.4845711	0.4617150	0.2740493

Sample sizes were computed using MARSSIM methodology for the maximum computed σ's for Ra-226 and Cs-137 in Tables 1 and 2, with α=0.01, β=0.01, and Δ=1 (Ra-226) and Δ=0.113 (Cs-137). These are presented in Table 3.

Table 3. Sample Size based on Maximum Sigma (σ) Computed from Historic Background Reference Area Data
Based on MARSSIM Table 5.3

		Off-site Lab Measurements		**On-site Lab Measurements**	
PARCEL	**SITEDSC**	**Cs-137**	**Ra-226**	**Cs-137**	**Ra-226**
PARCEL B	**Building 116**	21	25	15	21

Recommendations: Following MARSSIM guidance, an equal number of samples should be collected from the designated background area and the on-site SU. Sample sizes should be conservative and protective to human health and therefore be based on the greatest expected levels of variability. Sample size computations based on historical background reference area support the Navy's recommendation made on page 4-14 in Section 4.2.1 *Soil Background Reference Areas*, which is to collect a minimum of 25 samples per SU and background reference area. However as stated earlier, 25 samples should be collected per background reference area at surface and another 25 at depth, not across the five reference areas. This will result in 125 background reference area surface soil samples and 125 background reference area cores to be sampled at designated intervals.

Comment 6: Wilcoxon-Rank Sum Test (WRS)

It is unclear as to whether WRS tests will be performed to support remedial efforts. MARSSIM guidance clearly states that the WRS is to be used when ROCs are present in background surface soils. Historical sampling at the HPNS site confirms that RA-226 and Cs-137 can be found in background reference areas. However Section 6.6.2 of the Work Plan, Statistical Evaluation, states:

"*The statistical test presented in this Work Plan compares each analytical result for each ROC to the release criterion added to the mean for the background reference data set."*

This is a point-to-action level (AL) comparison not a population distribution comparison. Figure 6-2, Group 1 Soil Data Evaluation Process, indicates that the first step in data evaluation is to "Perform the $DCGL_w$ test". Again, as defined within the Work Plan this is a point-to-AL comparison. MARSSIM guidance and NRC guidance clearly indicate that comparisons of individual measurements to actionable levels is insufficient in determining whether or not a site meets the release criterion. Nuclear Regulatory Commission Publication NUREG-1505 refers to elevated measurement comparisons (EMC) which is similar to the methodology proposed by the Navy. NUREG-1505 states:

"*The EMC is intended to flag potential failures in the remediation process, and cannot be used to determine whether or not a site meets the release criterion until further investigation is done."*

There is statistical justification for requiring the WRS test or another test which accounts for variability and distributional characteristics of the sample data. Statistical tests such as the WRS test are hypothesis tests which are based on statistical inference. Statistical inference permits one to generate a conclusion about population characteristics based on information provided by a sample collected from that population. It provides a means for comparing the characteristics of one population sample to another population sample. In other words, the conclusions drawn from these tests can be applied to all of the un-sampled components of the population. There are also specified statistical levels of confidence associated with these tests. The proposed $DCGL_w$ test is only applicable to the single sample measurement that is being compared. The conclusion cannot be extrapolated to the remaining population (e.g., surface soil within an SU), and therefore cannot be used to determine if the release criterion has been met for an SU.

The hypotheses associated with the WRS test are:

> Null hypothesis (H_0): The median concentration in the SU exceeds that in the background reference area by more than the DCGL
> Alternate hypothesis (H_A): The median concentration in the SU exceeds that in the background reference area by less than the DCGL

It is possible for samples collected within an SU to exceed the release criterion, even if the final conclusion based on the WRS test is that the SU meets the release criterion. Because of the possibility of the presence of a few elevated concentrations, MARSSIM does recognize the need to support release/remedial efforts by comparing elevated measurements to the release criterion. However this is done in addition to the WRS test not instead of the WRS test. As stated earlier, results of those comparisons cannot be extrapolated to soils beyond where the discrete samples were collected with any statistical confidence.

It is incorrect to compare an individual sample measurement to a population parameter such as the mean in place of the WRS test. A possible alternative to computing the WRS test per SU, is to compute upper tolerance limits (UTLS) or upper prediction limits (UPLS) based on background reference data to which the individual sample measurements collected within an SU are compared. The UTL or UPL would become a background threshold value (BTV). This would provide a level of confidence associated with the comparisons. However, sample size calculations need to be based on the computation of these limits, not on the WRS test, and this method is not recommended when greater than six measurements will be compared (U.S. EPA, 2016).

Alternately, a UTL can be computed per SU and compared to a specified AL, such as a release criterion. Using a previously computed release criterion for Ra-226 at the HPNS site, 2.4 pCi/g, exploratory computations were performed to determine the sample size required for the computation of a non-parametric UTL using Visual Sample Plan (VSP) software. For at least 95% confidence that 95% of the population of surface soil within an SU has Ra-226 measurements below the AL, 59 samples would need to be collected. The non-parametric UTL was chosen to parallel the non-parametric choice of the WRS test by MARSSIM.

Recommendations: A minimum of 25 samples should be collected from appropriate background reference areas at appropriate depths and from each SU. It is recommended that the WRS test be used to support release of the individual

SUs, followed by comparison of the individual SU measurements to the appropriate release criterion to identify localized areas of high-level Ra-226 or Cs-137 contamination for possible remediation.

DISCUSSION

At the time of this review, the Work Plan presented by the Navy for assessment of the HPNS site is inconsistent in the discussed protocols for evaluating and interpreting the data that will be collected. This reviewer is in concurrence with the findings of the Navy's third party reviewer, that the procedures outlined in the current version of the Work Plan will provide insufficient data to support release of the HPNS Parcels. Although the Work Plan cites MARSSIM as guidance for the sample size determinations and the handling of background reference areas, the information provided in these Sections of the Work Plan do not always follow MARSSIM recommendations.

Additionally, comparisons of ROC measurements between on-site SUs and background reference areas are only addressed for Ra-226 and Cs-137. These ROCs are expected to be found in background. Other ROCs include, plutonium-239, strontium-90, thorium-232, and uranium-235. It is unclear from the Work Plan why these additional ROCs will not be compared to the project release criteria identified in Table 4-2 on page 4-12. MARSSIM provides guidance on applying the one-sample Sign Test for ROCs not found in background. Clarification regarding the evaluation of these ROCs is required before a review can be conducted.

Because of the allegations of fraud associated with historical data, the reliability of historical data is unknown at this time. Sample size calculations are driven by estimated variability and if the variability within the on-site SUs prove to be much greater than the variability of the historical data for the background reference areas, then appropriate statistical confidence and power will not be achieved in the WRS testing. A dynamic approach to designing survey/sampling activities would be the most defensible approach for the HPNS, with sampling activities broken down into phases. At the conclusion of each phase assumptions regarding the statistical distributions of the ROCs would be verified, and sample sizes adjusted, if needed.

REFERENCES

Nuclear Regulatory Commission (NRC). 1998. *Nonparametric Statistical Methodology for the Design and Analysis of Final Status Decommissioning Surveys*. Interim Draft Report for Comment and Use. U.S. Nuclear Regulatory Commission Office of Nuclear Regulatory Research. NUREG-1505, Rev. 1. https://www.nrc.gov/docs/ML0618/ML061870462.pdf (accessed on 2/18/2018)

MARSSIM. 2002. *Multi-Agency Radiation Survey and Site Investigation Manual (Revision 1)*. Nuclear Regulatory Commission NUREG-1575 Rev. 1, Environmental Protection Agency EPA 402-R-97-016 Rev. 1, Department of Energy DOE EH-0624 Rev. 1, August. https://www.epa.gov/radiation/multi-agency-radiation-survey-and-site-investigation-manual-marssim (accessed March 12, 2018).

MARSAME. 2009. *Multi-Agency Radiation Survey and Assessment of Materials and Equipment* (MARSAME). NUREG-1575, Supp. 1 EPA 402-R-09-001 DOE/HS-0004 https://www.epa.gov/radiation/marsame-manual-and-resources (accessed March 12, 2018).

U.S. Environmental Protection Agency. 2015. ProUCL 5.1 User Guide. Statistical Software for Environmental Applications for Data Sets with and without Nondetect Observations. EPA/600/R-07/041. https://www.epa.gov/land-research/proucl-version-5100-documentation-downloads

Visual Sample Plan (VSP). 2018. Pacific Northwest Laboratory. Version 7.1. https://vsp.pnnl.gov

cc: Central File - WA # SERAS-106 (w/attachment)
Electronic File - I:/Archive/SERAS/106/D/TM/031418
Kevin Taylor, SERAS Program Manager (cover page only)